INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(54) Title: MATERIALS AND METHODS RELATING TO DISEASE DIAGNOSIS
(54) Titre: MATERIAUX ET PROCEDES SE RAPPORTANT AU DIAGNOSTIC DE MALADIE

(57) Abstract

The invention provides materials and methods relating to disease diagnosis. In particular, the invention provides a method of diagnosing diseases, such as cancers, by comparing specific patterns of gene expression characteristic of the disease at a nucleic acid or protein level. The invention provides novel methods for analysing the expression profiles characteristic of diseased cells, in order to determine specific diagnostic markers. Such determined diagnostic markers may be stored on, for example a database, and used in the diagnosis of diseases such as cancer.

(57) Abrégé

La présente invention concerne des matériaux et des procédés se rapportant au diagnostic de maladie. Elle concerne en particulier un procédé de diagnostic de maladies, telles que des cancers, qui consiste à comparer des formes spécifiques d'expression génique caractéristique de la maladie à un niveau d'acide nucléique ou de protéine. Elle concerne aussi des procédés d'analyse des profils d'expression caractéristique de cellules malades, de manière à déterminer des marqueurs spécifiques de diagnostic. Ces marqueurs déterminés peuvent être conservés, par exemple dans une base de données, et utilisés dans le diagnostic de maladies telles que le cancer.

(51) International Patent Classification 7 :	A2	(11) International Publication Number: WO 00/70340
G01N 33/48		(43) International Publication Date: 23 November 2000 (23.11.00)

(21) International Application Number: PCT/EP00/04265

(22) International Filing Date: 11 May 2000 (11.05.00)

(30) Priority Data:
60/134,356   14 May 1999 (14.05.99)   US

(71) Applicant (for all designated States except US): KAROLIN-SKA INNOVATIONS AB [SE/SE]; Tomtebodavägen 11 F, Solna, S–171 77 Stockholm (SE).

(72) Inventors; and
(75) Inventors/Applicants (for US only): FRANZEN, Bo [SE/SE]; Astra Arcus AB, Preclinical R & D, S–151 85 Södertälje (SE). HAGMAN, Anders [SE/SE]; Astra Arcus AB, Analytical and Pharmaceutical R & D, D–151 85 Södertälje (SE). AYODELE, Alaiya [NG/SE]; Cancer Center Karolinska, S–Stockholm (SE).

(74) Agents: CRIPPS, Joanna, E. et al.; Mewburn Ellis, York House, 23 Kingsway, London WC2B 6HP (GB).

(81) Designated States: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published
Without international search report and to be republished upon receipt of that report.

(54) Title: MATERIALS AND METHODS RELATING TO DISEASE DIAGNOSIS

(57) Abstract

The invention provides materials and methods relating to disease diagnosis. In particular, the invention provides a method of diagnosing diseases, such as cancers, by comparing specific patterns of gene expression characteristic of the disease at a nucleic acid or protein level. The invention provides novel methods for analysing the expression profiles characteristic of diseased cells, in order to determine specific diagnostic markers. Such determined diagnostic markers may be stored on, for example a database, and used in the diagnosis of diseases such as cancer.

D scription

5

10

15

20

25

30

35

40

45

50

55

Materials and Methods Relating to Disease Diagnosis

Field of the Invention

The present invention concerns materials and methods
relating to disease diagnosis. Particularly, but not
exclusively, the invention relates to methods of
diagnosing tumours, by comparing specific patterns of
gene expression at a nucleic acid or protein level using
expressed nucleic acid, e.g. mRNA or cellular proteins
associated with the tumour.

Background of the Invention

The major characteristics that differentiate
malignant tumours from benign ones are their properties
of invasiveness and spread. Malignant tumours do not
remain localised and encapsulated: they invade
surrounding tissues, get into the body's circulatory
system, and set up areas of proliferation away from the
site of their original appearance. When tumour cells
spread and engender secondary areas of growth, the
process is call metastasis; malignant cells having the
ability to metastasize.

The earliest stages of malignant tumours are hard to
identify and pathologists are rarely sure how or where a
malignancy began. The cells of malignant tumours have a
tendency to lose differentiated traits and therefore it
can be difficult to determine the primary origin of the

- 2 -

cells following metastasis.

A concern with the histopathologic assessment of
neoplasias (tumour growth) is that tumour classification
is based on subjective evaluation (1, 2). Immunostaining
5   can be used to determine the expression of various
diagnostic markers and may increase reproducibility.
Ovarian cancer is an example of a disease where the
diagnostic difficulties are considerable (3). Epithelial
neoplasias of ovarian cancers are classified into benign,
10 borderline and malignant tumours. Borderline tumours are
often difficult to diagnose, and it is not known if most
of these tumours represent intermediate steps in tumour
progression or whether these tumours should be considered
as a separate group (4). Relative survival decreases with
15 increasing tumour stage or grade. Five-year survival is
considerably lower for women with carcinoma (38%) than
for women with borderline carcinoma (95%).

Summary of the Invention

20 The present inventors have appreciated that carrying
out routine tumour diagnosis in an accurate and objective
manner is very difficult. The process is preoperatively
dependent on an experienced cytologist and/or
postoperatively dependent on an experienced pathologist,
25 and is at present based on morphological judgements.
Further, the primary tumour source can be difficult to
determine which may lead to miss-diagnosis and

- 3 -

inappropriate treatment regime. Therefore, the present
inventors have realised that there is a need for a
diagnostic tool that can perform preoperative diagnosis
objectively. Such a tool should help to reduce the

5 number of patients undergoing unnecessary and expensive
therapy.

Multivariate analysis of the expression of a series
of diagnostic markers is one approach to diagnostic
problems. If a sufficiently large data set is collected,

10 it may be possible to recognize patterns of expression in
different histological groups. Goldschmidt et al. (5)
showed that multivariate analysis of 47 histological
variables generated by computer-assisted microscope
analysis facilitated classification of adipose tumours.

15 Similarly, multivariate analysis of RNA expression data
has been used to discriminate between fibroblast subtypes
(6).

One approach to obtain a large data set is to use
high resolution two-dimensional polyacrylamide gel

20 electrophoresis (2-DE). This technique is able to resolve
more than one thousand polypeptides on a single gel. The
pattern can be analysed by computer software such as
PDQUEST and MELANIE II (7, 8). This approach has been
previously used for the classification of lung tumour

25 cell lines (9).

An alternative approach to obtaining a large data
set is to use micro-array technology. Nucleic acid

- 4 -

sequence characteristic of nucleic acid sequences

expressed in certain cell types, e.g. MRNA or cDNA, can

be analysed in this way. There is an increasing tendency

towards miniaturisation of assays which use binding

5   members (such as antibodies or nucleic acid sequences).

For example, the binding members may be immoblised in

small discrete locations (microspots) and/or as arrays

(micro-array technology) on solid supports or on

diagnostic chips. These approaches can be particularly

10   valuable as they can provide great sensitivity

(particularly through the use of fluorescent labelled

reagents), require only very small amounts if biological

sample from individuals being tested and allow a variety

of separate assays to be carried out simultaneously.

15   Examples of techniques enabling miniaturised technology

are provided in WO84/01031, WO88/1058, WO89/01157,

WO93/8472, WO95/18376, WO95/18377, WO95/24649 AND EP-A-

0373203.

     Early research by Fedor et al established that

20   silicon could serve as a substrate onto which organic

molecules such as DNA could be synthesized. The process

now commercialised by Affymetrix Inc. Santa Clara,

California, involves the use of serial photolithographic

steps to build oligonucleotides in situ at a specific

25   addressable position on the chip.

     The strategy of addressing specific nucleic acid

sequences synthesized off chip, then hybridized to a

- 5 -

particular location on a chip by electrical attraction to
a charged microelectrode has been developed by Nanogen
Inc. Variation on the theme of microaddressable arrays
has recently led to the evaluation of chips for sequence
5   analysis of uncharaterised genetic material, mutational
analysis of a known gene locus, and for the evaluation of
a particular cell or tissue's profile of gene expression
for the whole complement of the human DNA sequence. These
methodologies typically relay on the use of laser
10   activated fluorescence of addressable signals on a
microchip.

      Thus, at its most general, the present invention
provides materials and methods for, firstly obtaining a
number of protein or nucleic acid expression profiles
15   characteristic for disease states of different origins or
different stages of development or malignancy; secondly,
analysing said expression profiles in order to determine
specific diagnostic markers; and thirdly, diagnosing the
presence of a disease, e.g. tumour, the type of disease
20   or the stage of development of said disease e.g. tumour
malignancy by comparison of its protein or nucleic acid
expression profile with those previously obtained to
determine using the specified diagnostic markers.

      Thus, the present invention primarily relates to a
25   method of obtaining gene expression profiles in order to
determine diagnostic markers characteristic of a selected
disease type or stage of development of a disease

- 6 -

comprising

     (1) obtaining cells from a sample of said disease tissue;

     (2) disrupting cells to expose the cellular products characteristic of gene expression;

     (3) separating said cellular products according to their characteristic properties on a substrate; and

     (4) carrying out computer-assisted multivariate analysis of the substrate to quantify and characterise the cellular product distribution on the substrate to identify specific diagnostic markers characteristic of said disease.

     Depending on the cell type, different genes are expressed or are expressed at different levels or frequency. These differences in gene expression may be used to characterise the type of cell. The cellular products that reflect the differences in gene expression are those products produced downstream of the nucleic acid transcription and translation process, e.g. mRNA or the expressed protein itself. These cellular products may then be separated according to their own characteristic properties, e.g. size, charge or sequence.

     In a preferred embodiment of the invention, the cellular products are expressed proteins which may be separated according to their size on a electrophoresis gel, preferably a two dimensional electrophoresis gel.

     Alternatively, the cellular products may be

- 7 -

separated according to their characteristic properties
using a substrate comprising specific binding members,
for example, antibodies or oligonucleotides. As mentioned
above, this is conveniently done by using a micro-array.

5   In such a situation, it is preferable to label the
cellular products, e.g. radioactively or fluorescently or
enzymatically, to assist in the computer-assisted
multivariate analysis.

Therefore, in a first aspect, the present invention

10   provides a method of obtaining protein expression
profiles in order to determine diagnostic markers
characteristic of selected disease types or stages of
disease development comprising

       (1) obtaining cells from a sample of said disease

15   type;

       (2) disrupting cells to expose the cellular proteins
contained therein;

       (3) separating said cellular proteins using a two-
dimensional electrophoresis gel; and

20   (4) carrying out computer-assisted multivariate
analysis of the two-dimensional electrophoresis gel to
quantify and characterise the protein distribution on the
gel to identify specific diagnostic markers
characteristic of said disease.

25   In order to carry out the analysis as outline in
step (4), quantitative and qualitative data from the two-
dimensional electrophoresis gel is firstly obtained.

- 8 -

Thus, step (4) may require carrying out multivariate
analysis of the quantitative and qualitative data from
the two-dimensional gel to characterise the protein
expression profile and identify specific diagnostic
5   markers characteristic of said disease.

In an alternative first aspect of the present
invention, there is provided a method of obtaining gene
expression profiles in order to determine diagnostic
markers characteristic of selected disease types or
10   stages of disease development, said method comprising

(1) obtaining cells from a sample of said disease
type

(2) disrupting cells to obtain the expressed nucleic
acid contained therein;

15   (3) separating said expressed nucleic acid using a
micro-array; and

(4) carrying out computer-assisted multivariate
analysis of the micro-array to quantify and characterise
the expressed nucleic acid on the micro-array to identify
20   specific diagnostic markers.

The expressed nucleic acid is preferably mRNA which
may be obtained from the cells by standard molecular
techniques known to the skilled person, for example see
Sambrook, Fritsch and Maniatis, "Molecular Cloning, A
25   Laboratory Manual", Cold Spring Harbor Laboratory Press,
1989, and Ausubel et al, Short Protocols in Molecular
Biology, John Wiley and Sons, 1992). Alternatively, cDNA

- 9 -

may be created from the expressed mRNA by reverse
transcription before separation and analysing on the
micro-array. Micro-array technologies use
oligonucleotides (representing thousands of different

5   genes) bound to given positions on various substrate.
Total mRNA is purified from a cell/tissue sample and cDNA
is produced by reverse transcriptase. Various steps (e.g.
in vitro transcription using biotinylated nucleotides)
may then be added before hybridisation and visualisation

10   depending on the specific type of micro-array technology
used (e.g. Affymetrix chips, Clontech membranes). The
final read-out is a signal that is proportional to the
quantity of a given expressed gene.

The present inventors have discovered that proteins

15   are differently expressed or differentially regulated
between various malignant tumours and benign tumours.

Therefore, the inventors believe that the present
invention will have particular utility in relation to the
diagnosis of tumours. Although the following description

20   of the invention concentrates on the diagnosis of tumours
in general, it will be appreciated by the skilled person
that the present invention may equally and advantageously
be applied to the diagnosis of other disease states
characterised by gene expression profiles, e.g.

25   hypo/hyperthyroidism, diabetes, or organ rejection.
Further, the invention may be used to test plasma samples
for leukaemia or other hematopoetic disorders.

- 10 -

In previous studies carried out by the present
inventors, a large degree of heterogeneity in protein
expression was observed, particularly in malignant
tumours (17, 18). Both qualitative and quantitative
5   differences were found within each tumour group.
However, the large quantitative variability indicated
that identification based on pattern recognition would be
difficult. However, the present inventors show herein
that it is possible to select a subset of variables which
10   show a characteristic pattern within the group, and thus
are useful for prediction of the presence of malignant
cells and their initial origin.

Thus, in a second aspect of the present invention,
there is provided a method of creating a collection of
15   diagnostic markers based on protein expression levels for
use in classifying disease cells in a given sample,
comprising

(1) obtaining cells from a plurality of samples of a
selected disease type;

20   (2) disrupting the cells to expose the cellular
proteins contained therein;

(3) separating the cellular proteins according to
their size on a two-dimensional electrophoresis gel for
each of said plurality of samples or a selected disease
25   type; and

(4) scanning said two-dimensional electrophoresis
gels to collect image data for each of the plurality of

- 11 -

samples of a selected disease type;

(5) analysing said image data in order to identify one or more markers characteristic of said selected disease type.

In an alternative second aspect of the present invention, there is provided a method of creating a collection of diagnostic markers based on nucleic acid expression levels for use in classifying disease cells in a given sample, comprising

(1) obtaining cells from a plurality of samples of a selected disease type

(2) disrupting the cells to obtain the expressed nucleic acid sequences contained therein,

(3) separating the expressed nucleic acids sequence according to their nucleotide sequence using micro-array technology for each of said plurality of samples of a selected disease type;

(4) scanning said micro-array to collect image data for each of the plurality of samples of a selected disease type; and

(5) analysing said image data in order to identify one or more markers characteristic of said selected disease type.

Again, the disease type is preferably cancer, wherein a plurality of samples may be collected from tumours of a particular cancer, e.g. ovarian, breast, skin etc, and its gene expression profile characterised

- 12 -

by the present invention.

It is important that the scanning of the
electrophoresis gel or the micro-array easily identifies
the separated proteins or nucleic acids respectively.

5   Therefore, the method may further comprise the step of
labelling the obtained proteins or expressed nucleic
acids. Nucleic acid sequences may be labelled by
standard techniques known to the skilled person such as
fluorescent, enzyme or radio-active labelling. As an

10   alternative to labelling obtained proteins, the gels may
be stained with, for example silver nitrate, and scanned
using a laser densitometer. Alternatively, the gels may
be analysed using computer-assisted microscope to
facilitate classification. The data obtained and

15   statistical comparison may be performed. In particular,
this is preferably a multivariate characterisation of one
or more numerical parameters associated with the
proteins. In other words, multivariate analysis of a
plurality of variables generated by, for example,

20   computer-assisted image analysis may be easily
classified. The statistical comparison may, for example,
identify a sub-set of proteins, from among all of the
proteins on the 2-DE, having a statistically significant
degree of expression and/or correlation when compared to

25   other samples from similar tumour cells. This sub-set of
proteins may then be used as diagnostic markers for the
particular tumour or stage of malignancy. Preferably, a

- 13 -

plurality of 2-DE gels are analysed and the distribution
pattern of the proteins are determined. A model may then
be set up with a specified number of variables between
the tumour cells being analysed. For example, a

5   comparison may be made between
benign/borderline/malignant. Preferably the number of
variables separating the groups whether proteins or
expressed nucleic acid sequences, will range between 20
and 500, more preferably 50 and 300, even more preferably

10   100 and 200. In general, it is preferably that the
number of variables is at least 20, more preferably at
least 50 and even more preferably at least 70, 100 or 150
variables. In the present case, the inventors used 170
variables.

15   Quantification and multivariate characterisation of
the expression profiles of selected protein or nucleic
acid groups may be performed on image analytical data
obtained from analysis of the 2-DE or the micro-array
respectively and used for objective classification of the

20   tumour cells in a given sample. The multivariate
characterisation may be carried out by partial least
squares discriminant analysis (PLS-DA). This process
allows (i) the construction and characterisation of a
protein or nucleic acid expression profile database and

25   data extraction of a plurality of sets of proteins or
nucleic acids which contribute significantly to the
diagnosis/classification of a disease state; (ii) add

- 14 -

samples/protein or nucleic acid expression profiles to

the database and further improve the future accuracy of

the diagnosis/classification; and (iii) query the

database via the expert system using new tumour

5   samples/protein or nucleic acid expression patterns

aiming at a prediction of diagnosis.

A protein expression profile database comprising

image data which has been analysed in order to determine

a plurality of variables for use as diagnostic markers;

10   said data being obtained from analysis of two-dimensional

electrophoresis gels showing characteristic protein

distribution associated with a disease type or state of

development of said disease for use in disease diagnosis

forms another aspect of the present invention.

15   A nucleic acid (mRNA or cDNA) expression profile

database comprising image data which has been analysed in

order to determine a plurality of variables for use as

diagnostic markers; said data being obtained from

analysis of a micro-array showing characteristic

20   expressed nucleic acid sequence distribution associated

with a disease type or stage of development of said

disease, for use in disease diagnosis forms yet another

aspect of the present invention.

In a further aspect, the present invention provides

25   a method of determining the presence, type or stage of a

disease type in a patient comprising the steps of

(1) extracting a sample of candidate disease cells

- 15 -

from the patient;

(2) disrupting the cells so as to expose the
cellular proteins contained therein;

(3) separating said cellular proteins on a two-
5   dimensional electrophoresis gel; and

(4) analysing said gel by computer assisted image
evaluation so as to compare protein distribution on gel
with a database of diagnostic markers characteristic of a
plurality of disease types or stages of disease
10 development to determine presence, type or risk of said
disease in said patient.

The present invention also provides a method of
determining the presence, type or stage of a disease in a
patient comprising the steps of

15 (1) extracting a sample of candidate disease cells
from a patient;

(2) disrupting the cells so as to obtain the
expressed nucleic acid sequences contained therein;

(3) separating said expressed nucleic acid sequences
20 on a micro-array according to their nucleotide sequence;
and

(4) analysing said gel by computer assisted image
evaluation so as to compare expressed nucleic acid
distribution on said micro-array with a database of
25 diagnostic markers characteristic of a plurality of
disease types or stages of disease development to
determine presence, type or risk of said disease in said

- 16 -

patient.

Preferably, the disease type is cancer and the disease cells are tumour cells.

Sample preparation may be carried out using standard techniques. One typical sample may contain approximately one million cells. Samples may be collected using fine needles aspiration biopsy (FNA) - a routine technique used for cytological diagnosis. The major advantage of using FNA combined with the expert system is (i) early diagnosis if possible, a prerequisite for making early decisions on therapy (ii) effects of hormone - or chemotherapy can be followed at protein expression level, providing early information on e.g. resistance against treatment; and (iii) the analysis is based on an average expression profile of the cell population.

Samples may also be collected after surgery for analysis in order to guide pathological examination and selection of post-operation therapeutic strategy.

As mentioned above, the earliest stages of malignant tumours are hard to identify and pathologists are rarely sure how or where a malignancy began. The present invention therefore has further utility in being able to more accurately determine the primary origin of tumour cells as the primary tumour and its corresponding metastasis express very similar 2-DE protein profiles (Franzen et al, Int. J. Cancer 1996, 69, 408-414). Such analysis will therefore assist a clinician in determining

- 17 -

the location of the primary tumour.

The above disclosure concentrates on the analysis and diagnosis of tumours. However, as mentioned above, the present invention may also be usefully applied to the diagnosis of any disease state that can be characterised by a statistically significant protein expression profile which allows the identification of specific diagnostic markers.

By way of example only, a brief outline/workflow on how the computer analysis may be set up in practice is provided below:

1. A new tumour sample is prepared, analyzed by 2-DE and the expression pattern is scanned.

2. All protein spots in this expression pattern is quantified and matched against a reference pattern using any established software for basic 2-DE analysis (e.g. PDQuest, Melanie, BioImage).

3. The data is first organized in a Excel-spreadsheet-like format table with all protein spot reference numbers in the first column and individual normalized protein quantities for every analyzed sample in the following columns. A new case/pattern is added as a new column. This corresponds to the "data table X".

- 18 -

4. The process of "data mining" - to find those
    proteins/variables which contribute most to the
    separation of tumour classes - and build the
    learning set (the core of the database), is based on
    5   the PLS-DA analysis. Here, an additional "data
    table Y" is included, as described under materials
    and methods, data preprocessing (please see also
    references 14 & 15). Graphically and numerically it
    is possible to make a first selection of variables
    10 (those that are far from origo (compare fig. 4) in
    the same and opposite direction from the
    corresponding position of tumour classes, compare
    fig. 3).

    15 5. In an interactive sub-routine or process, this first
    set of variables is crossvalidated by excluding
    cases one by one in sequences, rebuild the model and
    make a prediction of each of the excluded cases.
    Then, a second set of variables are selected
    20 (according to step 4), and so on - until the
    predictive value reach an optimum. In the present
    case, a set of 170 variables was selected in this
    way (step 4 and 5) and is therefore not a random
    choice.
    25

    6. Next, the true predictive value is determined using
    a new set of cases (the test set).

- 19 -

7. This process, step 3-6, can then be repeated with an
    increased number of cases in order to further
    improve the predictive accuracy.

8. A new case (an unknown tumour sample) is then
    analyzed by 2-DE/basic image analysis, the pattern
    is compared with respect to the defined group of
    variables in the database model and classified
    using, for example, PLS-DA prediction in order to
    obtain a diagnosis. Each new case may also be added
    to the database for future improvements of the
    predictive value of the model.

    One part of the expert system/computer software is
    to integrate steps 3 to 7 and make the process user-
    friendly in order to guide the investigator towards the
    construction of a model within the data base which
    provide high predictive accuracy. The other part of the
    expert system/computer software is to facilitate the
    query of the model using a new case in order to obtain a
    diagnosis (step 8 above). In addition to these
    "calculation parts" of the expert system, information may
    be included on sample preparation and on sample
    characteristics, 5-year survival data etc.

    Thus, in the further aspect of the present
    invention, there is a provided a diagnostic kit for
    diagnosing the presence, type or stage of a disease, e.g.

- 20 -

a tumour or malignancy of a tumour, said kit comprising a
database capable of quantifying an protein or nucleic
acid expression pattern and comparing it against
reference patterns held within the database. The kit may

5   also optionally include, instructions for carrying out
any of the methods described above; apparatus for
carrying out a 2-DE; micro-array technology or a laser
densitometer or other image scanning device.

    Aspects and embodiments of the present invention

10   will now be illustrated, by way of example, with
reference to the accompanying figures. Further aspects
and embodiments will be apparent to those skilled in the
art. All documents mentioned in this text are
incorporated herein by reference.

15

Brief Description of the Drawings

    Fig. 1   The two first principal components scores
(t_2 against t_1) of the 2-DE training data-set (22 gels and
1553 spots). A = benign ovary tumour sample (open

20   circles), B = borderline ovary tumour sample (mixed
circles), and C = malignant ovary tumour sample (filled
circles).

    Fig. 2   The two first principal components scores
(t_2 against t_1) of the most informative part of the 2-DE

25   training data-set (22 gels and 170 spots). For
descriptions, see Fig 1.

    Fig. 3   The two first PLS-DA scores (tPS_2 against

- 21 -

tPS$_1$) of the entire 2-DE data (40 gels and 170 spots). The
samples in the learning-set are indicated using circles (
A = benign ovary tumour sample (open circles), B =
borderline ovary tumour sample (mixed circles), and C =
5   malignant ovary tumour sample (filled circles). The
samples in the test-set are indicated using filled/mixed
and open squares in analogy with the learning-set.

     Fig. 4   The corresponding loading plot to Fig. 3
(wc$_2$ against wc$_1$). Indicated are the loading scores for
10   the most significant spots for separation of the three
tumour classes.

     Fig. 5 The two first principal components scores (t$_2$
against t$_1$) of breast tumour samples (33 gels and 170
spots). Cases classified as carcinoma are labelled "C"
15   and have filled symbols; cases classified as fibroadenoma
are marked with FA and have open symbols.

Detailed Description

20   1) MATERIALS AND METHODS

Tumour tissue samples

     All samples were obtained within 40 min after
resection and tumour cells were enriched as previously
25   described (10). Histopathological characterization was
carried out using hematoxylin-eosin stained sections of
formalin fixed and paraffin embedded specimens. Tumours

- 22 -

were classified using the WHO system.

Electrophoresis, scanning and image analysis

2-DE was performed as previously described (11).
5   Resolyte (2%, pH 4 - 8, BDH) were used for isoelectric
focussing, 10 - 13% linear gradient SDS-polyacrylamide gels
were used in the second dimension. Gels were stained with
silver nitrate as described by Rabilloud et al. (12) and
scanned at 100 mm resolution using a Molecular Dynamics
10   laser densitometer. Data was analysed using PDQUEST™
software (7) obtained from Pharmacia Biotech (Uppsala,
Sweden).

Data preprocessing

15   The data from the matchset was exported from PDQUEST
gel analysis package in the form of tables, with rows
representing gels and columns representing spots (data
table X - see references 14 and 15). Before the analysis,
the data was standardized by dividing each variable (table
20   column) by its standard deviation, thereby giving each
variable the same influence in the analysis. Thereafter the
data is centred by subtracting from each .column its
average.

Data analysis

25   The preprocessed data table (data table X) was
analysed by two data analysis methods. The first one,

- 23 -

Principal Component Analysis (PCA), extracts the
information in the data, in form of eigenvectors or
principal components. Visually, one can see this as a cloud
of points (the individuals cases/gels) in a
5   multidimensional space (each axis's representing each
spot). PCA first centers the data. Secondly, it rotates the
data in such a way that the greatest amount of linear
variation is described by the first component axis, the
residual variation is described by the second component
10   axis, and so on. Most of the information is often
compressed into two or three components. A more detailed
description of PCA may be found elsewhere (13).

The second data analysis method, Partial Least
Squarest (PLS) - Discriminant analysis, was used to
15   classify the cases into the three tumour-classes (benign,
borderline or malignant). An additional data table (data
table Y) with the classification of the tumours is included
into the analysis. Table Y consists of the same number of
columns as the number of tumour classes and the number of
20   rows is equal to the number of cases. The table is then
filled with suitable dummy variables (i.e. 1 = belongs to
a specific tumour class or 0 = does not belong).

The PLS-analysis is similar to PCA in that it projects
the data table X into a vector. It differs, however, in
25   that the direction of the vector is determined both by the
variation of data table X (as in the case of PCA) as well
as the variation of data table Y. For further descriptions

- 24 -

of PLS, see (14, 15). The significance of the PLS-model is checked by cross-validation. Data from a small number of samples is kept out of the calculation, the PLS model is computed from the remaining data, and the y-values of the

5   deleted are thereafter predicted from the model. The differences in square between predicted and actual y-values for deleted samples are summed to form PRESS (Predictive Error of Sum Squares). This sequence is repeated until each sample has been deleted once.

10   The data-table used for training the PLS-model consists of 22 cases and 170 spots (Table X). To test the model a table (18 cases and 170 spots) with unknown tumour class was used (Table X).

    The data analysis were carried out on CODEX™ software

15   obtained from Sumit System AB (Stockholm, Sweden) and SIMCA™ software obtained from Umetri AB (Umeå, Sweden).

2) RESULTS

20   <u>Creation of a Learning Set</u>

    Cells were extracted from fresh ovarian tumour tissue and single cell suspensions free of erythrocytes were prepared (11). Cytological smears were prepared from all preparations and samples usually contained > 90% tumour

25   cells (histopathological characteristics are presented in Table 1). 2-DE polypeptide patterns obtained from these cells were analysed by the PDQUEST™ software (7). The

- 25 -

patterns of polypeptide expression in 22 ovarian tumours
were examined, 5 benign (A), 6 borderline (B) and 11
malignant (C) cases (objects). These patterns were matched
together and a reference 2-DE map was constructed
5   containing 1553 spots (variables).

As an initial step, principal component analysis was
applied to entire material (22 gels and 1553 spots) to
provide an overview over the data structure, to identify
outliers and possible clusters. Normalized quantities
10   (expressed as ppm) for all spots were used for the PCA.
Fig. 1 shows the scores for the first two components. A
coarse separation into two major groups, A + B and C was
observed, indicating that latent structures with predictive
value are present in this set of data. However, the
15   corresponding loading plots showed very scattered data
(data not shown).

Of the original data (1553 variables, Fig. 1), 170
variables had a substantial influence on the model (PLS
loadings > 0.02). Approx. 100 variables were active in
20   separating the groups A + B (benign/borderline) and C
(malignant), and approximately 70 variables in separating
between A (benign) and B (borderline). An improved
separation of the clusters representing each of the three
classes was observed using these 170 variables (Fig. 2).
25   Four significant PLS-DA vectors were found, by using cross-
validation (Q2=0.84), describing 98.4 % of the variance in
Y and 40.7 % in X. This data set was then closed and

- 26 -

called "learning set".

Testing the model with unknown tumours

Eighteen new cases were analysed by 2-DE and added to the existing matchset. Expression levels of the 170 markers for all cases were analysed blindly using PCA, enabling the distribution of new objects. Figure 3 show the predictions of unknown cases in a PLS score plot (and the corresponding loadings in Fig. 4).

After breaking the code, 6 of 8 malignant cases were correctly classified. Case 84 and 89 were classified as borderline. Furthermore, 3 of 4 borderline cases were correctly classified, whereas borderline case 96 was classified as benign. Benign cases 90 and 95 were correctly were correctly classified. Of the remaining 4 cases, 3 were classified as borderline and one (case 29) as borderline/malignant.

Testing a ovary model with breast tumours

The possibility that an ovarian cancer model could be used for classification intraductal breast tumours was exploited. The present inventors matched the ovary tumour matchset standard 2-DE map with a corresponding breast tumour standard map in the database (16). Seventy-five of the 170 markers were present in the breast standard map. Fig. 5 shows the PCA distribution of 33 cases of breast cancer (26 carcinomas, 6 fibroadenomas and 1 normal breast

- 27 -

epithelium). Only a tendency of clustering of benign cases
was observed which indicate that some but not all of the
markers show predictive value.

5   3) DISCUSSION

      The present inventors present here a first attempt to
apply artificial learning strategies using quantitative 2-
dimensional electrophoresis data for tumour diagnosis. A
subset of the information in the 2-DE pattern, based on 170

10   spots, was selected. Using these variables, a learning set
was constructed where an acceptable separation of the
groups benign/borderline/malignant tumours into three
clusters was obtained. Whether other combinations of spots
will result in an improved separation is unknown and

15   difficult to test, since each learning set has to be tested
by a new panel of unknown samples. We tested the learning
set using 18 cases, and observed a correct classification
of the majority of these (11/18).

      It is well known among pathologists that the routinely

20   used limited number of diagnostic sections may not be
representative for a certain lesion. In this context it is
important to note that the sampling technique employed for
2-DE analysis is more likely to meet the requirements for
lesion representivity.

25   In previous studies by the present inventors, a large
degree of heterogeneity in polypeptide expression was
observed, particularly in malignant tumors (17, 18). Both

- 28 -

qualitative and quantitative differences were found within each tumour group. Particularly, the large quantitative variability indicated that identification based on pattern recognition would be difficult. The present data suggests

5   that it is possible to select a subset of variables which show limited variability within the group, and useful for prediction.

Neural networks and artificial learning has been used to predict cancer prognosis and for grading tumors (5, 19-

10   22). The parameters used have been various TNM-scoring systems, nuclear grading, tumour markers and histopathological scoring. For prostate cancer, the sensitivity of the network was between 81 to 100% and the specificity 72 to 75% to predict various outcomes such as

15   seminal vesicle and lymph node involvement (22). Similarly, neural network analysis has been performed on breast cancer, using parameters such as hormone receptor status, DNA index, tumour size, number of axillary lymph nodes involved with tumour as input information (20). These

20   studies have indicated that artificial learning is a powerful method to increase the diagnostic accuracy on individual tumours.

The present inventors have noted that many of the alterations observed in 2-DE pattern are similar between

25   tumours of epithelial origin. Thus similar changes in the expression of some cytoskeletal and stress proteins are observed in breast, ovarian and prostate tumors (10; Alaiya

- 29 -

et al., unpublished). With this background, it was interesting to examine whether a selected set of ovarian markers could be used for classification of intraductal breast tumors into benign and malignant. Some clustering of

5   benign cases was observed, whereas malignant cases showed extensive scattering. It seems reasonable to suggest that it will be difficult to construct a universal model for epithelial tumors, and that learning sets have to be created for each tumour type.

10   In conclusion, the present study suggests that artificial learning strategies can be used for tumour diagnosis.

- 30 -

REFERENCES

1.   Dalton, L. W., Page, D. L., and Dupont, W. D.., Cancer.
73: 2765-2770, 1994.

2.   Kronqvist, P., Montironi, R., Kuopio, T., and Collan,
Y. U. Anal. & Quant. Cytology & Histology. *19:* 423-429,
1997.

3.   Friedlander, M. L., Seminars in Oncology. *25:* 305-314,
1998.

4.   Link, C. J. J., Reed, E., Sarosy, G., and Kohn, E. C.,
Am. J. Medicine. *101:* 217-225, 1996.

5.   Goldschmidt, D., Decaestecker, C., Berthe, J. V.,
Gordower, L., Remmelink, M., Danguy, A., Pasteels, J. L.,
Salmon, I., and Kiss, R., Lab. Invest. *75:* 295-306, 1996.

6.   Spanakis, E. and Brouty-Boye, D. Int. J. Cancer. *71:*
402-409, 1997.

7.   Garrels, J. I., J. Biol. Chem. *264:* 5269-5282, 1989.

8.   Wilkins, M. R., Hochstrasser, D. F., Sanchez, J. C.,
Bairoch, A., and Appel, R. D., Trends in Biochem. Sciences.
21: 496-497, 1996.

9.   Schmid, H. R., Schmitter, D., Blum, P., Miller, M., and
Vonderschmitt, D., Electrophoresis. *16:* 1961-1968, 1995.

10.   Alaiya, A. A., Franzén, B., Fujioka, K., Moberger, B.,
Schedvins, K., Silversvård, C., Linder, S., and Auer, G.,
Int. J. Cancer. *73:* 678-683, 1997.

- 31 -

11. Franzén, B., Okuzawa, K., Linder, S., Kato, H., and
Auer, G., Electrophoresis. *14:* 382-390, 1993.

12. Rabilloud, T., Vuillard, L., Gilly, C., and Lawrence,
J.-J.   A general overview., Cell Mol. Biol. *40:* 57-75,
1994.

13. Jocliffe, I. T. New York: Springer Verlag, 1986.

14. Jellum, E., Harboe, M., Bjune G., Wold S. J. Pharm. &
Biomedical Analysis 9: 663-669, 1991.

15. Hagberg, G. A review of pattern recognition methods.
NMR in Biomedicine. *11:* 148-56, 1998.

16. Franzén, B., Auer, G., Alaiya, A. A., Eriksson, E.,
Uryu, K., Hirano, T., Okuzawa, K., and Linder, S. down-
regulation of cytokeratins, Br. J. Cancer. *73:* 1632-1638,
1996.

17. Franzén, B., Auer, G., Alaiya, A. A., Eriksson, E.,
Uryu, K., Hirano, T., Okuzawa, K., Kato, H., and Linder, S.
Int. J. Cancer. *69:* 408-414, 1996.

18. Alaiya, A. A., Franzén, B., Linder, S., and Auer, G.,
Electrophoresis. in press.

19. Dawson, A. E., Austin, R. E. J., and Weinberg, D. S.,
J. Clinical Pathology. *95:* 29-37, 1991.

20. Ravdin, P. M., Clark, G. M., Hilsenbeck, S. G., Owens,
M. A., Vendely, P., Pandian, M. R., and McGuire, W. L.,
Breast Cancer Res. & Treat. *21:* 47-53, 1992.

21. Erler, B. S., Hsu, L., Truong, H. M., Petrovic, L. M.,
Kim, S. S., Huh, M. H., Ferrell, L. D., Thung, S. N.,

5

- 32 -

Geller, S. A., and Marchevsky, A. M., Lab. Invest. *71:* 446-

451, 1994.

22.   Tewari, A. and Narayan, P., J. of Urology. *160:* 430-

436, 1998.

5

Claims

5

10

15

20

25

30

35

40

45

50

55

Claims

1.   A method of obtaining combinations of gene
expression profiles in order to determine diagnostic
markers characteristic of a selected disease type or
stage of development of a disease comprising
        (1) obtaining cells from a sample of said disease
tissue;
        (2) disrupting cells to expose the cellular products
characteristic of gene expression;
        (3) separating said cellular products according to
their characteristic properties on a substrate; and
        (4) carrying out computer-assisted multivariate
analysis of the substrate to quantify and characterise
the cellular product distribution on the substrate to
identify specific diagnostic markers characteristic of
said disease.

2.   A method according to claim 1 wherein the cellular
products characteristic of gene expression are proteins.

3.   A method according to claim 1 or claim 2 wherein the
substrate is an electrophoresis gel which allows
separation of the cellular products characteristic of
gene expression according to their size.

4.   A method according to claim 3 wherein said gel is
2D-electrophoresis gel.

5.   A method according to claim 1 wherein the cellular
products characteristic of gene expression are nucleic
acid sequences.

6.   A method according to claim 5 wherein the nucleic
acid sequences are mRNA.

7.   A method according to claim 1, claim 5 or claim 6

34

wherein the substrate comprises a plurality of binding
members capable of binding said cellular products
characteristic of gene expression.

5   8.   A method according to claim 7 wherein said binding
members are oligonucleotides capable of binding said
cellular products characteristic of gene expression
according to their nucleotide sequence.

10   9.   A method according to claim 1 or claim 2 wherein
said binding members are antibodies.

       10.   A method according to any one of claims 7 to 9
wherein is said substrate is a micro-array.

15   11.   A method according to any one of the preceding
claims wherein said cellular products characteristic of
gene expression are labelled to assist computer-assisted
multivariate analysis.

20   12.   A method according to any one of the preceding
claims wherein said multivariate analysis is carried out
by partial least squares discriminant analysis (PLS-DA).

25   13.   A method according to any one of the preceding
claims wherein the disease is cancer and the cells are
tumour cells or normal reference cells within a given
tumour.

30   14.   A method of creating a collection of diagnostic
markers based on protein expression levels for use in
classifying disease cells in a given sample, comprising
       (1) obtaining cells from a plurality of samples of a
selected disease;
35   (2) disrupting the cells to expose the cellular
proteins contained therein;
          (3) separating the cellular proteins on a two-

35

dimensional electrophoresis gel for each of said
plurality of samples of the selected disease; and
        (4)   scanning said two dimensional electrophoresis
gels to collect image data for each of the plurality of
5   samples of the selected disease.

15.   A method of creating a collection of diagnostic
markers based on nucleic acid expression levels for use
in classifying disease cells in a given sample,
10   comprising
        (1) obtaining cells from a plurality of samples of a
selected disease;
        (2) disrupting the cells to obtain the expressed
nucleic acid sequences contained therein;
15   (3) separating the expressed nucleic acid sequences
on a micro-array for each of said plurality of samples of
the selected disease; and
        (4)   scanning said micro-array to collect image data
for each of the plurality of samples of the selected
20   disease.

16.   A method according to claim 14 or claim 15 further
comprising the step of analysing said image data in order
to identify one or more markers characteristic of said
25   selected disease.

17.   A method of determining the presence, type or stage
of a disease in a patient comprising the steps of
        (1) extracting a sample of candidate disease cells
30   from the patient;
        (2) disrupting the cells so as to expose the
cellular proteins contained therein;
        (3) separating the cellular proteins on a two-
dimensional electrophoresis gel; and
35   (4) analysing said gel by computer assisted image
evaluation so as to compare protein distribution on gel
with a database of diagnostic markers characteristic of a

36

plurality of tumour types or stages of malignancy to determine presence, type or risk of said disease in said patient.

18. A method of determining the presence, type or stage of a disease in a patient comprising the steps of
    (1) extracting a sample of candidate disease cells from the patient;
    (2) disrupting the cells so as to obtain the expressed nucleic acid sequences contained therein;
    (3) separating the expressed nucleic acid sequences on a micro-array according to their individual nucleotide sequence; and
    (4) analysing said micro-array by computer assisted image evaluation so as to compare expressed nucleic acid distribution on said micro-array with a database of diagnostic markers characteristic of a plurality of disease types or stages of development of said disease to determine presence, type or risk of said disease in said patient.

19. A method according to any one of the preceding claims wherein the number of markers characteristic of said disease type is in the range of 20 to 500.

20. A method according to claim 19 wherein the number of markers characteristic of said disease type is in the range of 50 to 300.

21. A method according to any one of claims 14 to 20 wherein the disease type is selected from the group cancer, hypo/hyperthyroidism, diabetes, organ rejection, and samples for leukaemia or other hematopoetic disorders.

22. A method according to claim 21 wherein said disease state is cancer and said disease tissue is a tumour.

37

23. A protein expression profile database comprising image data which has been analysed in order to determine a plurality of variables for use as diagnostic markers; said data being obtained from analysis of two dimensional
5   electrophoresis gels showing characteristic protein distribution associated with disease type and state of disease for use in disease diagnosis.

24. A protein expression profile database according to
10 claim 23 wherein said disease is cancer and the state of said diseases equates to the state of malignancy of said cancer.

25. A nucleic acid expression profile database
15 comprising image data which has been analysed in order to determine a plurality of variables for use as diagnostic markers; said data being obtained from analysis of a micro-array showing characteristic expressed nucleic acid distribution associated with disease type and state of
20 disease for use in disease diagnosis.

26. A nucleic acid expression profile database according to claim 25 wherein said disease is cancer and the state of said diseases equates to the state of malignancy of
25 said cancer.

27. A nucleic acid expression profile database according to claim 25 or claim 26 wherein the expressed nucleic acid is mRNA or cDNA.
30

Table I: Histopathological characteristics of samples

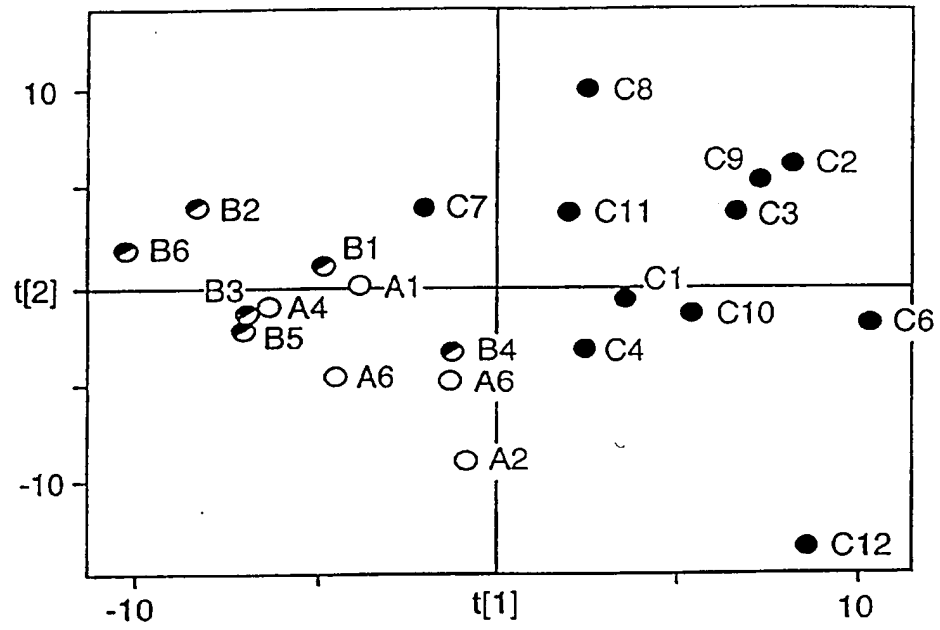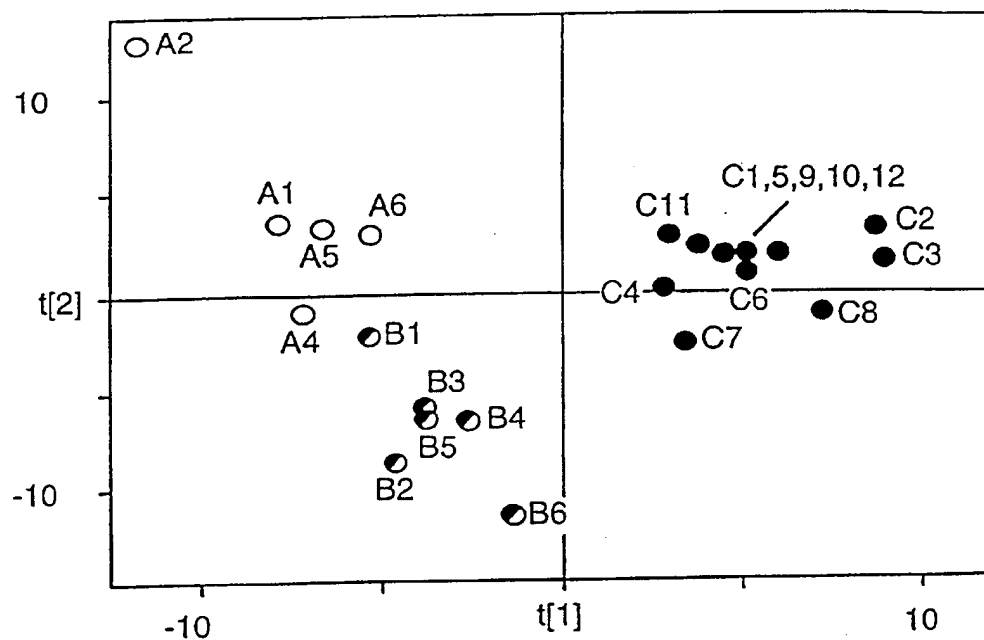| Serial No.	Case-No.	Learning model label	Test Cases: Predicted Result from PLS-DA	True Type (A/B/C)	Pathological Diagnosis
1	OC14	A1		A	Serous Cystadenoma IA
2	OC19	A2		A	Serous Cystadenoma IA
3	OC34	A4		A	Serous Cystadenoma IA
4	OC38	A5		A	Serous Cystadenoma IA
5	OC26	A6		A	Mucinous Cystadenoma IIA
6	OC82		B	A	Cystadenofibroma
7	OC39	B1		B	Borderline Seropapillary IB
8	OC46	B2		B	Borderline Seropapillary IB
9	OC50	B3		B	Borderline Seropapillary IB
10	OC21	B4		B	Borderline Mucinous IIB
11	OC59	B5		B	Borderline Mucinous IIB
12	OC68	B6		B	Borderline Mucinous IIB
13	OC72		B	B	Borderline Serous
14	OC77		B	B	Borderline Serous
15	OC07	C1		C	Sero Papillary ADC(IC)
16	OC08	C2		C	Sero Papillary ADC(IC)
17	OC09	C3		C	Sero Papillary ADC(IC)
18	OC20	C4		C	Seropapillary IC
19	OC30	C6		C	Bil Seropapillary IC
20	OC40	C7		C	Bil Adenocarcinoma
21	OC43	C8		C	Bil Seropapillary IC
22	OC04	C9		C	Mixed tumor
23	OC06	C10		C	Clear Cell tumor (IVC)
24	OC27	C11		C	Clear Cell tumor (IVC)
25	OC33	C12		C	Endometrioid Ca IIIC
26	OC48		C	C	Sero Papillary IC
27	OC45		C	C	Endometrioid Ca IIIC
28	OC90		A	A	Serous Cystadenofibroma
29	OC96		A	B	Borderline Serous
30	OC49		C	C	Endometrioid Ca IIIC
31	OC84		B	C	Clear Cell tumor (IVC)
32	OC74		C	C	Endometrioid Ca IIIC
33	OC73		C	C	Sero Papillary ADC(IC)
34	OC89		B/C	C	Sero Papillary ADC(IC)
35	OC95		A	A	Serous Cystadenoma IA
36	OC29		B	A	Mucinous Cystadenoma IIA
37	OC66		B	A	Serous Cystadenoma IA
38	OC35		B	A	Serous Cystadenoma IA
39	OC111		C	C	Sero Papillary ADC(IC)
40	OC117		B	B	Borderline Mucinous IIB

FIG. 1

FIG. 2

FIG. 3

FIG. 4

FIG. 5